ΘΕΜΑΤΑ ΔΙΠΛΩΜΑΤΙΚΩΝ ΕΡΓΑΣΙΩΝ

Εργαστήριο Συστημάτων Βάσεων Γνώσεων & Δεδομένων

GoNToggle: ΕΞΥΠΝΗ ΜΗΧΑΝΗ ΑΝΑΖΗΤΗΣΗΣ ΜΕ ΧΡΗΣΗ ΟΝΤΟΛΟΓΙΩΝ

ΠΛΗΡΟΦΟΡΙΕΣ: Θ. Δαλαμάγκας, 210 7721402, dalamag@dblab.ntua.gr, Σ. Σουλδάτος, 210 7721402, stef@dblab.ntua.gr

ΠΕΡΙΛΗΨΗ: Η διπλωματική εργασία στοχεύει στον σχεδιασμό μιας έξυπνης μηχανής αναζήτησης που θα χρησιμοποιεί οντολογίες. Η μηχανή θα μπορεί να χρησιμοποιηθεί για την αναζήτηση κειμένων διαφόρων μορφών (π.χ. doc, ppt, pdf, txt, ps, κ.λ.π.) σε προσωπικούς υπολογιστές, συλλογές κειμένων εταιρειών, κ.λ.π.

ΑΤΟΜΑ: 1-2

ΠΛΑΤΦΟΡΜΑ ΕΡΓΑΣΙΑΣ: C++/Java

ΣΥΝΤΟΜΗ ΠΕΡΙΓΡΑΦΗ: Οι μηχανές αναζήτησης είναι δημοφιλές εργαλείο εύρεσης πληροφοριών στον Ιστό. Αρκούν κάποιες λέξεις κλειδιά π.χ. στο διάσημο Google για να βρεί κάποιος ιστοσελίδες και αρχεία σχετικά με το θέμα του. Ειδικότερα, το scholar google ειδικεύεται στην αναζήτηση κειμένων χρήσιμων σε σπουδαστές (π.χ. επιστημονικές εργασίες, αναφορές, κ.λ.π.). Η τεχνολογία της μηχανής αναζήτησης Google, μπορεί να χρησιμοποιηθεί ήδη και ως αυτόνομο πρόγραμμα αναζήτησης αρχείων σε προσωπικούς υπολογιστές (http://desktop.google.com/). Τα προγράμματα αυτά λέγονται *desktop search engines*.
Η ιδέα πίσω από τέτοια εργαλεία είναι ότι σαρώνουν τους φακέλους στον σκληρό δίσκο, μαντεύουν το είδος του αρχείου (doc, pdf, ps, κ.λ.π.), και χρησιμοποιούν το περιεχόμενο ή μέρος του αρχείου, δηλ. τις λέξεις, ώστε να χτίσουν κάποιας μορφής ευρετήριο (index). Όταν ο χρήστης διατυπώσει μια ερώτηση με χρήση λέξεων-κλειδιών, τότε το ευρετήριο βρίσκει ποια αρχεία έχουν αυτές τις λέξεις, και με βάση στατιστικά μοντέλα το σύστημα εκτιμά ποια από τα αρχεία έχουν περιεχόμενο σχετικό με τις λέξεις-κλειδιά του χρήστη.

Η διπλωματική εργασία στοχεύει στην σχεδίαση και στην υλοποίηση ενός *desktop search engine (GoNTogle)* για την αναζήτηση κειμένων διαφόρων μορφών (π.χ. doc, ppt, pdf, tct, ps, κ.λ.π.) σε προσωπικούς υπολογιστές, συλλογές κειμένων εταιρειών, κ.λ.π. Η διαφορά με τις υπάρχουσες *desktop search engines* είναι ότι το *GoNTogle* θα κάνει χρήση οντολογιών. Η αναζήτηση δε θα περιορίζεται στην χρήση λέξεων-κλειδιών, αλλά ο χρήστης θα μπορεί να αναζητά κείμενα που έχουν χαρακτηριστεί με κάποιες έννοιες. Για παράδειγμα, θα μπορεί να αναζητά κείμενα που έχουν χαρακτηριστεί ως ‘εργασίες μαθήματος’, χωρίς να είναι απαραίτητο να υπάρχει ο προσδιορισμός αυτός στο ίδιο το περιεχόμενο του αρχείου.
Για να πραγματοποιηθεί μια τέτοια αναζήτηση είναι αναγκαίες οντολογίες για διάφορα πεδία γνώσης. Ο χρήστης θα μπορεί να χρησιμοποιεί έτοιμες οντολογίες ή να κατασκευάζει καινούριες. Επίσης ο χρήστης θα πρέπει να αναθέτει στα κείμενά του έννοιες από αυτές τις οντολογίες είτε χειρωνακτικά είτε βοηθούμενος από το σύστημα το οποίο και θα μπορεί να του

προτείνει έννοιες σχετικές με το κείμενό του. Για τη συγκεκριμένη διπλωματική, οι οντολογίες θα ορίζονται με βάση το πλαίσιο RDFS (http://www.w3.org/TR/rdf-primer/).

Η αναζήτηση κειμένων στο GoNTogle θα μπορεί να γίνει επομένως με δύο τρόπους:

(α) Κλασσικός: αναζήτηση κειμένων με χρήση λέξεων-κλειδιών όπως σε ένα κλασσικό desktop search engine.

(β) Προχωρημένος: αναζήτηση κειμένων με χρήση οντολογιών. Ο χρήστης θα μπορεί να χρησιμοποιεί μια οντολογία, να πλοηγείται στις έννοιες και τις σχέσεις της, και το σύστημα να εμφανίζει κείμενα που έχουν χαρακτηριστεί σημασιολογικά με στοιχεία από την οντολογία αυτή. Επιπλέον, το σύστημα θα μπορεί να εκμεταλλεύεται την σημασιολογία της οντολογίας και να προτείνει έννοιες πιο στενές (ή πιο ευρείες) από τις αρχικές που θα μπορούν να χρησιμοποιηθούν για την αναζήτηση σε περίπτωση που τα αποτελέσματα είναι πάρα πολλά (ή πολύ λίγα) και δεν ικανοποιούν τον χρήστη. Π.χ. αν το σύστημα δεν βρίσκει κείμενα σχετικά με ‘τεχνολογίες αντικειμενοστρεφών βάσεων’ ειδικά, θα μπορεί να προτείνει κείμενα σχετικά με ‘τεχνολογίες βάσεων’ γενικότερα.

Η διπλωματική αυτή θα ασχοληθεί επομένως με τα παρακάτω θέματα:

1. Ανάλυση απαιτήσεων ενός desktop search engine GoNTogle για την αναζήτηση κειμένων διαφόρων μορφών (π.χ. doc, ppt, pdf, txt, ps, κ.λ.π.) σε προσωπικούς υπολογιστές, συλλογές κειμένων εταιρειών, κ.λ.π.
2. Σχεδίαση GoNTogle
3. Μελέτη διαχείρισης και ανάπτυξης οντολογιών RDFS.
4. Ανάπτυξη ευρετηρίων κειμένων με εξαγωγή πληροφοριών και μεταδεδομένων από κείμενα.
5. Σημασιολογικός χαρακτηρισμός (semantic annotation) κειμένων με βάση στοιχεία οντολογίας.
6. Υλοποίηση GoNTogle

Για περισσότερες πληροφορίες, δείτε την αναφορά στη σελίδα:
http://www.dbnet.ece.ntua.gr/~dalamag/pub/r3014.pdf